Title	漢日英Universal Dependencies平行コーパスとその差異
Author(s)	安岡, 孝一
Citation	じんもんこん2019論文集 (2019), 2019: 43-50
Issue Date	2019-12
URL	http://hdl.handle.net/2433/245218
Right	ここに掲載した著作物の利用に関する注意 本著作物の著作権は情報処理学会に帰属します。本著作物は著作権者である情報処理学会の許可のもとに掲載するものです。ご利用に当たっては「著作権法」ならびに「情報処理学会倫理綱領」に従うことをお願いいたします。; The copyright of this material is retained by the Information Processing Society of Japan (IPSJ). This material is published on this web site with the agreement of the author (s) and the IPSJ. Please be complied with Copyright Law of Japan and the Code of Ethics of the IPSJ if any users wish to reproduce, make derivative work, distribute or make available to the public any part or whole thereof. All Rights Reserved, Copyright ©2019 Information Processing Society of Japan.
Type	Conference Paper
Textversion	publisher

漢日英Universal Dependencies平行コーパスとその差異

安岡孝一 (京都大学人文科学研究所附属東アジア人文情報学研究センター)

言語横断的な依存構造記述である Universal Dependencies は，ニューラルネットを用いた言語解析ツール等に採用されており，言語をまたいだ係り受け解析に非常に有用である．しかしながら，Universal Dependencies にもとづくコーパスの開発は，各言語ごとに独立しておこなわれており，そこでは各言語固有の事情が優先されている，というのも，また事実である．では，各言語固有の事情によって，各言語の Universal Dependencies は，どの程度バラバラになってしまっているのか．本稿では，古典中国語 UD・近代日本語 UD・近代英語 UD による『大學』平行コーパスを作成した上で，その差異について述べる．

Universal Dependencies Parallel Corpora on Classical Chinese, Modern Japanese, and Modern English

Koichi Yasuoka (Kyoto University)

Universal Dependencies Treebank is a cross-linguistic project to annotate Parts-of-Speech and dependency relations universally. The same 17 universal Part-of-Speech tags and 37 universal dependency-relation tags are used for the annotation universally across all languages. However, in fact, each UD treebank is developed by each developer of each language, reflecting “not-universal” treatments of the language. In this paper, we reveal the difference among Classical Chinese UD, Modern Japanese UD, and Modern English UD, upon parallel corpora of 大學 (The Great Learning).

1 はじめに

筆者が班長を務める京都大学人文科学研究所共同研究班「東アジア古典コーパスの実証研究」(班員: ウィッテルン クリスティアン，守岡知彦，池田巧，山崎直樹，二階堂善弘，鈴木慎吾，師茂樹，李媛，白須裕之，藤田一乗) では，現在，古典中国語 (漢文) の依存文法解析に精力を傾注しており，その道具立ての一つとして，Universal Dependencies(以下「UD」)[1] の古典中国語への適用を研究している．依存文法解析それ自体は，Tesnière [2] の構造的統語論に源を発し，Мельчук [3] の有向グラフ記述によって，一応の完成を見た手法である．その最大の特長は，言語横断的な記述が可能だという点にあり，Мельчук の手法をコンピュータ向けに洗練した UD においても，言語に関わらない記述，という特長が前面に押し出されている．

ところが，カレル大学の UD プロジェクトに参画し，筆者らが製作した UD Classical Chinese-Kyoto 四書コーパスを UD 2.4 [4] へ含めるにあたって，この言語横断的な記述という点で，いくつかの齟齬が生じた．齟齬の多くは，単語への品詞付与で発生したが，どうやらその原因は，単語とその品詞に対する考え方が，彼らとわれわれの間で，形容しがたい谷を隔てているためだと感じられた．

ただ，形容しがたい，と言っていても始まらない．この形容しがたい谷を，何とか言語化して捕まえるべく，本稿では，多言語平行コーパス [5] を，UD で試作した．具体的には，漢籍の翻訳文を UD 化し，それらの間の比較をおこなうものであり，手始めに『大學』の古典中国語 UD に対し，日本語 UD と英語 UD による比較をおこなってみることにした．

2 古典中国語UD・近代日本語UD・近代英語UDによる平行コーパス

四書の中で最も短い『大學』(1753 字) の古典中国語 UD に対し，近代日本語の読み下し文による UD と，英訳文による UD を，平行コーパスとして作成した．

近代日本語 UD については，小牧昌業の読み下し文 [6] を原テキストとして，旧仮名口語 UniDic [7] と MeCab 0.996 [8] で形態素解析をおこない，UD Japanese-Modern [9, 10] にもとづいて品詞を変換し，UDPipe 1.2.0 [11] の japanese-gsd-ud-2.4 モデルで係

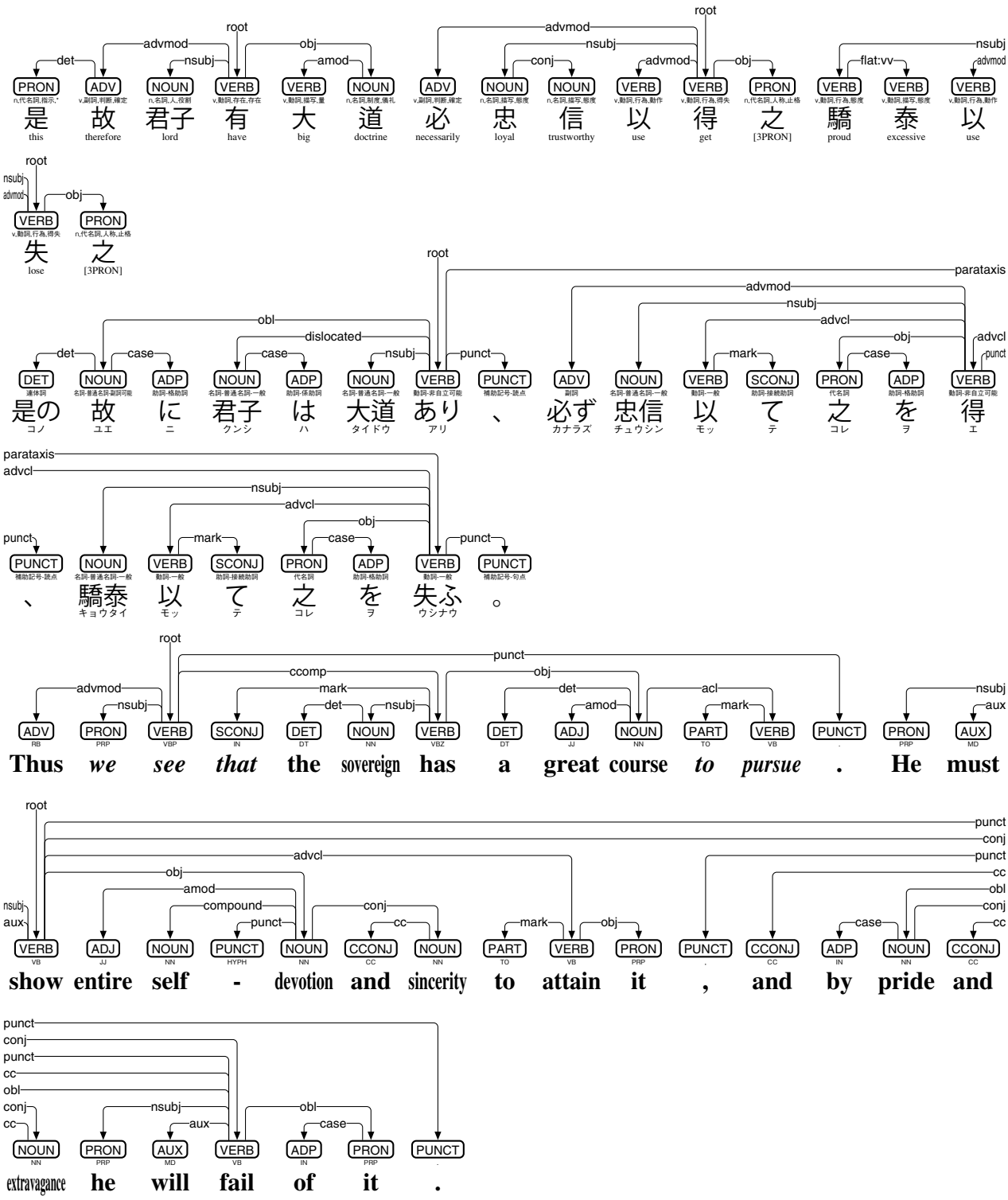

図 1: 「是故君子有大道必忠信以得之驕泰以失之」の漢日英 UD 平行コーパス

り受け解析をおこなった後に，解析ミスなどに対する修正[a]を手作業で施した．

近代英語 UD については，James Legge の英訳 [14] を原テキストとして，StanfordNLP 0.2.0 [15] の en_ewt モデル 0.2.0 で形態素解析・係り受け解析をおこなった後に，わずかな修正を手作業で施した．

これら漢日英 UD による『大學』平行コーパスに対し，語境界，品詞付与，係り受け，文境界，解釈，の 5 つの視点から，差異を比較検討した．これらの比較結果のうち，特に重要と考えられる差異を，以下に述べる．

2.1 語境界の差異

われわれの古典中国語 UD においては，固有名詞や一部の名詞を除いて，1 文字を 1 語としている．日本語 UD では，短単位 [16] を 1 語とみなしている [10]．英語 UD においては，わずかな例外 (たとえば「cannot」を「can」と「not」に分ける) を除いて，空白および句読点 (ハイフンや引用符を含む) が語境界である．

この結果，たとえば「忠信」は，古典中国語 UD においては 2 語，近代日本語 UD においては 1 語とみなされる．これに対応する「self-devotion and sincerity」は，近代英語 UD においては 5 語である (図 1)．語境界が対応せず，したがって，語そのものも 1 対 1 には対応しない．

『大學』平行コーパスにおける単語数 (句読点を含む) は，古典中国語 UD が 1729，近代日本語 UD が 2709，近代英語 UD が 3682 となっている (表 1)．近代日本語 UD の単語数が，古典中国語 UD より増えているのは，日本語固有の助詞や助動詞 (表 1 の ADP や AUX) が増えているためだと考えられる．また，これらに比べ，近代英語 UD における単語数がさらに多いのは，代名詞や冠詞 (表 1 の PRON や DET) が多いためだと考えられる．

2.2 品詞付与の差異

英語 UD における品詞は，語の機能に応じて付与される．たとえば「last」という語は，基本的に

表 1: 漢日英 UD『大學』平行コーパスの品詞分布

	漢	日	英
ADJ	-	88	196
ADP	28	567	302
ADV	125	53	153
AUX	33	198	301
CCONJ	55	86	117
DET	-	88	343
INTJ	3	2	4
NOUN	457	586	575
NUM	15	12	11
PART	154	10	118
PRON	159	71	408
PROPN	18	24	61
PUNCT	-	353	587
SCONJ	28	98	60
SYM	-	-	-
VERB	654	473	446
X	-	-	-
合計	1729	2709	3682

は ADJ(形容詞) だが，連用修飾をおこなう場合には ADV(副詞) とみなされる．

日本語 UD における品詞は，語の形態に応じて，UniDic 品詞体系を変換する形で付与される [10]．たとえば「終に」は，1 語 (ついに) の場合には ADV とみなされるが，「終」「に」の 2 語 (おわりに) の場合には NOUN(名詞) と ADP(後置詞・前置詞) とみなされる．

古典中国語 UD における品詞は，MeCab-Kanbun における品詞体系 [17] を変換する形で付与している [18]．たとえば「終」という語は，全て VERB(動詞) とみなす．連体修飾をおこなう場合も，やはり VERB とみなす．というのも，われわれは形容詞という品詞区分を設けておらず，動詞は連体修飾を (あるいは連用修飾も) おこなえる，という言語モデルを採用しているから[b]である．

[a] 現代日本語 UD ならば，たとえば赤塚忠の通釈 [12] を GiNZA 2.2.0 [13] で形態素解析・係り受け解析した方が，はるかに高精度の UD が得られる．しかし，著作権上の問題が残ることから，本稿では，旧字旧仮名の近代日本語 UD を用いることにした．なお，本稿で使用した解析ツール UniDic2UD は，python モジュールの形でパッケージ化をおこない，`https://pypi.org/project/unidic2ud` で公開した．

[b] この点に関しては，カレル大学のメンバーと大激論になった．彼らとしては，連体修飾語は ADJ，連用修飾語は ADV という品詞が，当然あたえられるものと考えていたようである．ただ，その一方で，名詞が連体修飾や連用修飾をおこなう場合は NOUN のままでいい，と彼らも考えており，その齟齬が気になった．結局，UD Classical Chinese-Kyoto 四書コーパス [4] においては，連体修飾や連用修飾をおこなう動詞は VERB のままとなったが，その語の形態素情報 (FEATS フィールド) に VerbForm=Part や VerbForm=Conv を記す，というのが落としどころとなった．古典中国語は屈折しないにもかかわらず，「連体屈折形」「連用屈折形」を記すことになったわけである．

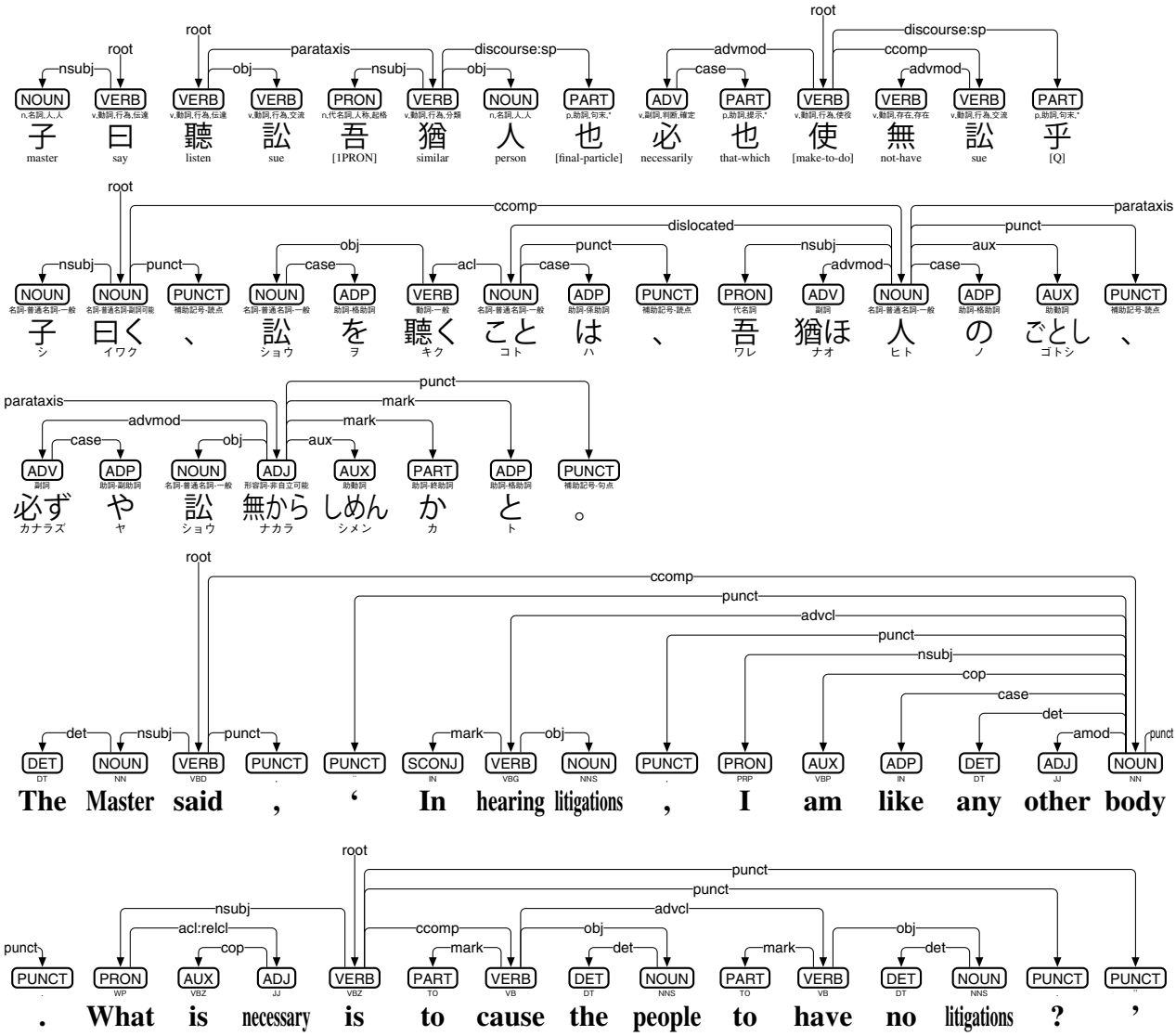

図 2: 「子曰聽訟吾猶人也必也使無訟乎」の漢日英 UD 平行コーパス

たとえば，古典中国語 UD における「大道」は,「大」が VERB,「道」が NOUN である．これに対応する「a great course」は，近代英語 UD においては,「a」が DET(冠詞・連体詞),「great」が ADJ,「course」が NOUN である．近代日本語 UD においては,「大道」は 1 語とみなされ，NOUN である (図 1).

あるいは「吾猶人也」の「猶」は，古典中国語 UD においては VERB である．これに対し，近代日本語 UD においては,「猶ほ」は ADV,「ごとし」は AUX(助動詞) である．近代英語 UD においては,「am like」の「am」は AUX [c],「like」は ADP である (図 2).

この結果,『大學』平行コーパスでは，近代英語 UD における ADJ 数と VERB 数の合計は，古典中国語 UD の VERB 数に少しだけ足りない (表 1)．近代日本語 UD における ADJ 数と VERB 数の合計は，さらに少なくなっている．

古典中国語 UD における「是」は PRON(代名詞) だが，近代英語 UD における「this」は，体言を修飾している場合は DET，そうでない場合は PRON である．近代日本語 UD においては,「この」と読むならば DET,「これ」と読むならば PRON とみなしている．

この結果,『大學』平行コーパスでは，近代日本語 UD における DET 数と PRON 数の合計は，古典中国語 UD の PRON 数と一致している (表 1)．一方，近代英語 UD の DET は冠詞を含んでおり，PRON は関係代名詞などを含んでいることから，いずれも大幅に増加している．

[c]近代英語 UD『大學』コーパスでは，214 ある「be」動詞のうち，AUX が 194 で，VERB はわずか 20 である．コピュラ文や受動文での「be」は，AUX とみなされることに注意が必要である．

表 2: 漢日英 UD『大學』平行コーパスの係り受け分布

	漢	日	英
acl	42	86	19
acl:relcl	-	-	52
advcl	28	260	102
advmod	173	95	183
amod	33	-	107
appos	-	-	10
aux	33	188	110
aux:pass	-	-	90
case	68	546	290
cc	47	0	117
ccomp	56	39	48
clf	3	-	-
compound	8	9	20
compound:prt	-	-	13
compound:redup	9	-	-
conj	71	-	112
cop	5	10	97
csubj	27	22	13
csubj:pass	-	-	1
dep	-	-	-
det	65	67	334
det:predet	-	-	7
discourse	8	5	4
discourse:sp	56	-	-
dislocated	1	8	0
expl	9	-	23
fixed	12	0	2
flat	12	-	2
flat:vv	19	-	-
goeswith	-	-	-
iobj	7	7	5
list	0	-	0
mark	31	171	143
nmod	44	29	114
nmod:npmod	-	-	2
nmod:poss	-	-	104
nsubj	224	222	274
nsubj:pass	0	-	77
nummod	15	12	9
obj	249	189	185
obl	19	197	158
obl:lmod	22	-	-
obl:tmod	22	-	-
orphan	0	-	2
parataxis	16	67	26
punct	-	353	587
reparandum	-	-	-
root	293	126	173
vocative	1	1	1
xcomp	1	-	66
合計	1729	2709	3682

2.3 係り受けの差異

漢日英『大學』平行コーパスにおける係り受けタグ 50 種類 (言語固有タグ 13 種類を含む) の分布を，表 2 に示す．各単語には，係り受けリンクが 1 本ずつ入っていることから，係り受け数の合計は，単語数の合計 (表 1) に等しくなる．

nsubj(主語) と obj(目的語) は，漢日英で係り受け数の差があり，しかも，その内容が非常に異なっている．たとえば，図 1 の「君子有大道」においては，古典中国語 UD では「君子」が nsubj,「大道」が obj であり，近代英語 UD でも同様だが，近代日本語 UD では「大道」が nsubj となってしまうため,「君子」は dislocated(外置語) 扱い[d] となっている．一方,「驕泰以失之」においては，古典中国語 UD では「驕泰」が nsubj,「之」が obj であり，近代日本語 UD でも同様だが，近代英語 UD では「pride and extravagance」も「it」も前置詞を伴って obl(斜格補語) となっており，さらに「he」が nsubj として追加されている．あるいは，図 3 の「詩云」においては，古典中国語 UD では「詩」は nsubj だが，近代日本語 UD では「詩」は obl となっている．近代英語 UD でも「the Book of Poetry」は obl となっており，さらに「it」が nsubj:pass(主語 [受動文]) として追加されている．

人名と称号との間の扱いは，漢日英で，かなり異なっている．図 3 の「師」と「尹」を見てみよう．古典中国語 UD では,「師」から「尹」へ flat(並列) を繋いでおり，名詞「師」が固有名詞「尹」の称号だとみなしている．近代日本語 UD では,「尹」から「師」へ nmod(体言による連体修飾語) を繋いでおり，固有名詞「尹」を名詞「師」が修飾している，という立場を取っている．近代英語 UD では,「teacher」から「Yin」へ appos(同格) を繋いでおり，これらが同格で交換可能である，という立場を取っている．

古典中国語 UD の係り受けタグのうち，特徴的なのは discourse:sp(談話要素 [文助詞]) である．古典中国語の動賓終構造 (predicate-object-final structure) において,「終」にあたる構造を記述するために導入 [19] したものである．具体的には，図 2 の「吾猶人也」の「也」や,「使無訟乎」の「乎」への係り受けリンクに使用しており，近代日本語 UD や近代英語 UD には対応する要素がない．

flat:vv(並列 [動詞類]) は，古典中国語 UD において，動詞連続を記述するために導入 [20] したものである．図 1 では，動詞の「驕」と「泰」を繋いでいる．近代英語 UD では，対応する「pride」と「extravagance」を conj(接続) で繋いでおり，近代日本語 UD では「驕泰」は 1 語である．ただし，古典中国語 UD の重畳語においては，compound:redup(複合 [重畳]) を導入しており，図 3 の「巖」と「巖」,「赫」と「赫」が典型例である．近代日本語 UD では「巖巖」も「赫赫」も 1 語であり，近代英語 UD では対応する要素が文全体に訳し込まれているようである．

近代英語 UD の係り受けタグのうち，特徴的なのは nmod:poss(体言による連体修飾語 [所有]) である．代名詞の所有格において，det(決定詞) を使用せず，あえて nmod:poss にしており，図 3 の「its rugged masses」の「its」が典型例である．その一方

[d] 日本語 UD では,「象は鼻が長い」という例文 [10] に対し,「鼻」を nsubj,「象」を dislocated としている．

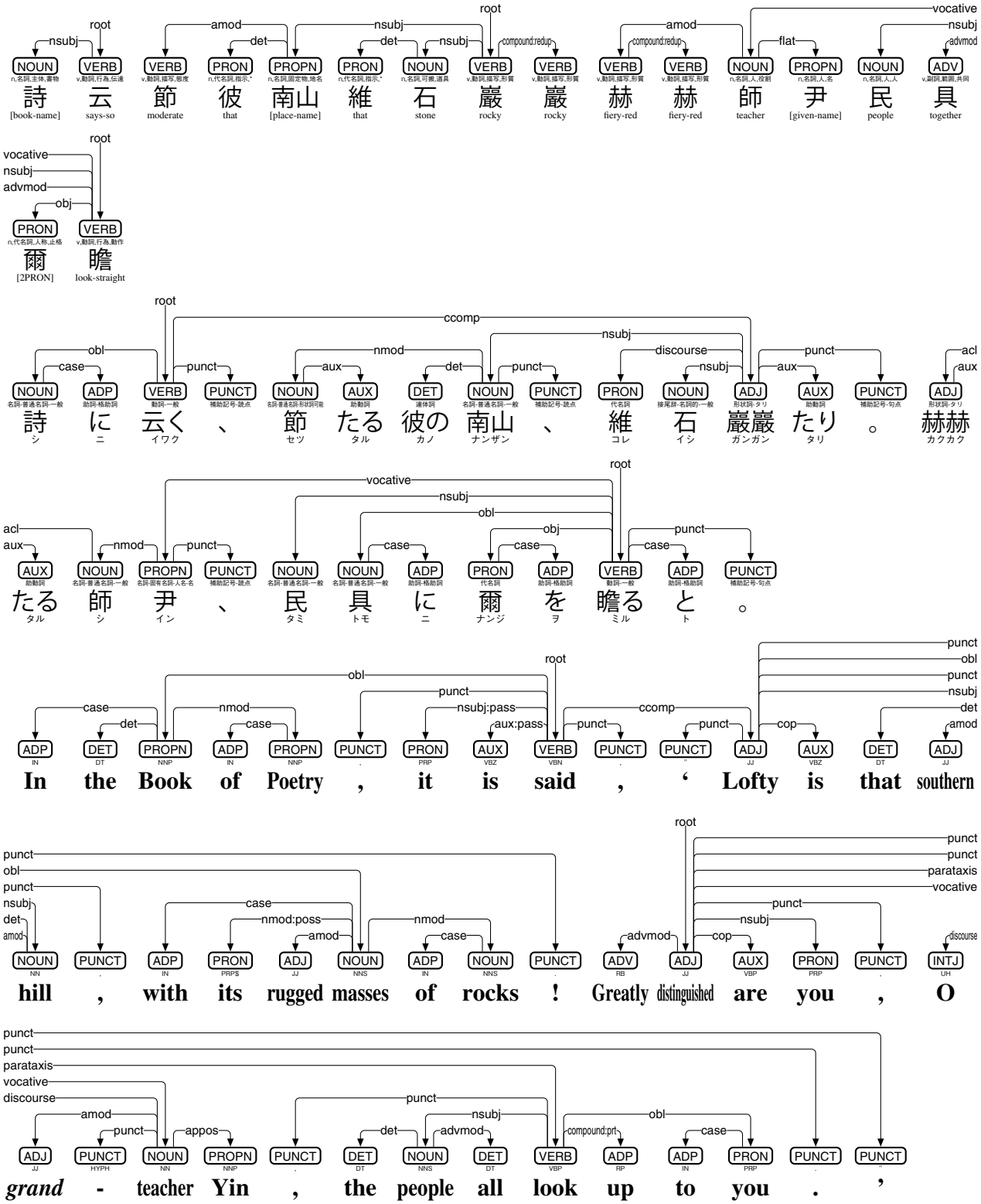

図 3: 「詩云節彼南山維石巖巖赫赫師尹民具爾瞻」の漢日英 UD 平行コーパス

で,「all」や「such」による連体修飾は，det を拡張して det:predet(決定詞 [前置]) としており，独特のこだわりが感じられる．

2.4 文境界の差異

古典中国語においては，文という単位は自明ではなく，連続する漢字列からどのように文を切り出すかは，かなり恣意的である．

たとえば，図 1 の「是故君子有大道必忠信以得之驕泰以失之」に対しては，われわれの古典中国語 UD においては，これを 3 つの文に分割している．一方，小牧昌業は，これを 1 文で読み下しており，それが近代日本語 UD に反映されている．あるいは James Legge は，これを 2 文に分割しており，それが近代英語 UD に反映されている．

近代英語 UD には引用符が含まれており，これが文境界を，さらに複雑にしている．図 2 の「The Master said, ‘ In hearing litigations, I am like any other body. What is neccessary is to cause the people to have no litigations ?’」[e)]では，StanfordNLP の en_ewt モデルは引用符内のピリオドを文境界とみなし，なおかつ，引用の前半を「The master said,」とくっつけてしまう．結果として，これを近代英語 UD は 2 文とみなしている．一方で，古典中国語 UD は「子曰聽訟吾猶人也必也使無訟乎」を 3 つの文に分割しており，近代日本語 UD は 1 文にまとめている．

あるいは，図 3 の例では，近代英語 UD と近代日本語 UD は 2 文，古典中国語 UD は 3 文である．この結果，漢日英『大學』平行コーパスにおいては，古典中国語 UD での文の数は 293，近代日本語 UD での文の数は 126，近代英語 UD での文の数は 173 となっている (表 2 の root)．これに加え，古典中国語 UD は『禮記』を底本としているが，近代日本語 UD と近代英語 UD は朱熹の入れ替え版に依っている．文の数が違う上に，順序まで入れ替わってしまっているため，対応づけが非常に複雑になってしまっている．

2.5 解釈の差異

古典中国語 UD においては，白文を MeCab-Kanbun で形態素解析した後，手作業で係り受け解析をおこなっており，そこには，われわれの解釈が入り込むことになる．近代英語 UD においては，英訳の時点で James Legge の解釈が入っており，その上に en_ewt モデルによる解釈が加わることになる．近代日本語 UD においては，読み下し文に小牧昌業の解釈が入っており，さらに，旧仮名口語 UniDic と UD Japanese-Modern と japanese-gsd-ud-2.4 による解釈が加わることになる．

例として，図 3 の「維」を見てみよう．古典中国語 UD では，この「維」は「石」を修飾するとみなして読んでいる．近代英語 UD でもほぼ同様で,「its」が「masses of rocks」を修飾している．ところが近代日本語 UD では,「維」は語気を整える助詞「これ」とみなして読んでいる．はっきり解釈が異なっているのだ．

当然ながら，これらの解釈が同一となることは有り得ない．というのも，これらの解釈は，元の白文における曖昧性を，それぞれに解消 (あるいは軽減) するためのものである．どの曖昧性に，どう着目し，どう解消するかは，それぞれの解釈で異なっているからである．

3 おわりに

Universal Dependencies は本当に Universal なのか，という問いに対しては，現状では否と答えざるを得ない．古典中国語 UD と近代日本語 UD と近代英語 UD を比較してみただけでも，様々なレベルでの差異があり，しかも，これらの差異が埋まることは無さそうである．UD 2.4 は，83 の言語にまたがっており，全体としての差異は，かなり凄いことになっているのは想像に難くない．

ただし，各言語ごとにバラバラにおこなわれてきた自然言語処理ツールの開発を，兎にも角にも一まとめにした，という点では UD の功績は大きい．UD によって，対照言語学や比較言語学は新たなツールを手に入れた，と考えることもできるだろう．本稿のような手法，すなわち複数の言語にまたがる差異の比較が可能となったのも，UD という比較ツールを，いわばモノサシの一つとしているわけである．

なお，本稿で作成した漢日英『大學』平行コーパスを，表 3 の URL で公開している．是非ダウンロードして，比較してみてほしい．

参考文献

[1] Joakim Nivre: Towards a Universal Grammar for Natural Language Processing, CICLing 2015: 16th International Conference on Intelligent Text

[e)] 引用符前後の空白は [14]p.364 に依った．

表 3: 漢日英『大學』平行コーパス公開 URL

古典中国語 UD	https://corpus.kanji.zinbun.kyoto-u.ac.jp/gitlab/Kanbun/ud-kanbun/tree/master/kanripo/kR1d0052/043
近代日本語 UD	https://corpus.kanji.zinbun.kyoto-u.ac.jp/gitlab/Kanbun/ud-ja-kanbun/tree/master/kokuyaku/A01A
近代英語 UD	https://corpus.kanji.zinbun.kyoto-u.ac.jp/gitlab/Kanbun/ud-en-kanbun/tree/master/Legge/THE_GREAT_LEARNING

Processing and Computational Linguistics (April 2015), pp.3-16.

[2] Lucien Tesnière: Éléments de Syntaxe Structurale, Paris: C. Klincksieck (1959).

[3] Igor A. Mel'čuk: Dependency Syntax: Theory and Practice, New York: State University of New York Press (1988).

[4] Joakim Nivre, Daniel Zeman, et al.: Universal Dependencies 2.4, LINDAT/CLARIN digital library at the Institute of Formal Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Prague: Charles University (May 2019).

[5] 山崎直樹: 多言語平行コーパスのための言語横断的な構造記述，2009 中日理論言語学国際フォーラム発表論文集 (2009 年 7 月)，pp.26-27.

[6] 國譯漢文大成，經子史部，第一卷 (1922 年 10 月)，國民文庫刊行會.

[7] 小木曽智信: 旧仮名遣いの口語文を対象とした形態素解析辞書，人文科学とコンピュータ「じんもんこん 2012」論文集 (2012 年 11 月)，pp.25-32.

[8] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoro: Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (July 2004), pp.230-237.

[9] Mai Omura, Yuta Takahashi, Masayuki Asahara: Universal Dependency for Modern Japanese, Proceedings of the 7th Conference of Japanese Association for Digital Humanities (September 2017), pp.34-36.

[10] 浅原正幸，金山博，宮尾祐介，田中貴秋，大村舞，村脇有吾，松本裕治: Universal Dependencies 日本語コーパス，自然言語処理，Vol.26，No.1 (2019 年 3 月)，pp.3-36.

[11] Milan Straka and Jana Straková: Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe, Proceedings of CoNLL 2017 Shared Task (August 2017), pp.88-99.

[12] 赤塚忠: 大学・中庸，新釈漢文大系，第 2 巻 (1967 年 5 月)，東京: 明治書院.

[13] 松田寛，大村舞，浅原正幸: 短単位品詞の用法曖昧性解決と依存関係ラベリングの同時学習，言語処理学会第 25 回年次大会発表論文集 (2019 年 3 月), pp.201-204.

[14] James Legge: The Chinese Classics, Second Edition, Vol.I, Oxford: Clarendon Press (1893).

[15] Peng Qi, Timothy Dozat, Yuhao Zhang, Christopher D. Manning: Universal Dependency Parsing from Scratch, Proceedings of the CoNLL 2018 Shared Task (October 2018), pp.160-170.

[16] 近藤明日子: 近代文語 UniDic 短単位規程集, Ver.1.1，立川: 国立国語研究所コーパス開発センター (2016 年 3 月).

[17] 安岡孝一，ウィッテルン クリスティアン，守岡知彦，池田巧，山崎直樹，二階堂善弘，鈴木慎吾，師茂樹: 古典中国語 (漢文) の形態素解析とその応用，情報処理学会論文誌，Vol.59，No.2 (2018 年 2 月)，pp.323-331.

[18] 安岡孝一: 古典中国語 Universal Dependencies で読む『孟子』, センター研究年報 2018 別冊，京都: 京都大学人文科学研究所附属東アジア人文情報学研究センター (2019 年 3 月).

[19] Herman Leung, Rafaël Poiret, Tak-sum Wong, Xinying Chen, Kim Gerdes and John Lee: Developing Universal Dependencies for Mandarin Chinese, 12th Workshop on Asian Language Resources (December 2016), pp.20-29.

[20] 安岡孝一: Universal Dependencies にもとづく古典中国語 (漢文) の依存文法解析, センター研究年報 2018, 京都: 京都大学人文科学研究所附属東アジア人文情報学研究センター (2018 年 10 月).

